2015

Evaluation of R Package ltm with IRT Dichotomous Models

Fusun Sahin
University of Albany - SUNY, fsahin@albany.edu

Kimberly Colvin
University of Albany - SUNY, kcolvin@albany.edu

Evaluation of R Package *ltm* with IRT Dichotomous Models

Füsun Şahin

Kimberly F. Colvin

University at Albany, SUNY

Abstract

There are many software packages that estimate item response theory parameters and examinee abilities. This study evaluates the accuracy of the item parameter and ability estimates generated by the open-source R package *ltm*. In this simulation study, item and ability estimates were compared to the true parameters under six conditions that differed in the numbers of items and examinees. After looking at the resulting bias, mean absolute deviation, and root mean square error, we concluded that item parameter and ability estimates from *ltm* were estimated reasonably accurately with results similar to previous studies of established commercial software.

Keywords: IRT estimation, dichotomous models, parameter estimation, R

Evaluation of R Package *ltm* with IRT Dichotomous Models

Study Purpose

There are many software packages for psychometric analyses and specifically for use with item response theory (IRT) models. The software packages available in the open-source R program are gaining attention, which even led to a special volume on use of R packages in psychometrics in the *Journal of Statistical Software* (JSS, 2007). Despite the abundant availability of such packages, a critical aspect of these software packages is the accuracy of estimations. Simulation studies are effective for evaluating the accuracy of estimations since such studies allow researchers to compare estimates with the true values.

Some recent simulation studies have studied the accuracy of IRT analyses conducted with open-source R package *ltm* (Rizopoulos, 2006, 2013) by comparing it to commercial programs such as BILOG-MG (Zimowski, Muraki, Mislevy and Bock, 2003) for analyzing either dichotomous or polytomous responses (Bulut & Zopluoglu, 2013; Pan, 2012). Previous studies generated simulation conditions by manipulating ability and item parameters and modeled the responses by using IRT models. The dichotomous responses were modeled by a one-parameter logistic model (1PL). This study contributes to the existing literature by extending the comparisons to two- and three-parameter logistic (2PL and 3PL) models under various conditions on the number of items and examinees, while keeping the ability and item parameter distributions similar with previous studies.

Theoretical Framework

The unidimensional IRT model constitutes the theoretical foundation of the study. It is assumed that the set of items measure only one underlying ability (or latent-trait), which affect

the probability of examinees' responses to individual items. Dichotomous responses were generated to represent either a correct or an incorrect response.

IRT models are functions of items, characterized by item parameters, and the ability of the examinees. Three item parameters used in IRT models are: difficulty, *b*; discrimination, *a*; and pseudo-guessing, *c* (Hambleton, Swaminathan, & Rogers, 1991). The 1PL or Rasch model (Rasch, 1960) is based only on item difficulty, the 2PL (Lord, 1952) uses both difficulty and discrimination, and the 3PL (Birnbaum, 1968) uses difficulty, discrimination, and pseudo-guessing parameters. All three of these IRT models also include ability (θ), representing the ability of an examinee on the latent-trait of interest.

The theoretical boundaries for each parameter are different. While in theory, the ability of an examinee can be considered on a scale from negative to positive infinity, in practice, ability is often quantified between -3 and +3. Since ability and item difficulty are on the same scale, item difficulty can take both negative and positive values. In practice, *b* values often range between -2 and 2 where smaller values indicate easier items. Although, theoretically, the discrimination parameter *a* can range from negative to positive infinity, in practice only items with positive *a* values are used. A negatively differentiating item means that for an examinee with lower ability there is a higher probability of providing a correct response to that item; therefore, items with negative *a* values are considered problematic and eliminated from tests. Moreover, it is also not usual to obtain *a* values great than two. Therefore, in practice *a* values range between 0 and 2 (Hambleton, et al, 1991). Third, *c* represents the probability of an examinee with infinitely low ability correctly answering the item (Hambleton et al.). Since *c* represents a probability, it ranges from 0 to 1, where larger *c* values indicate not well-written items.

Methodology

First, one free, open-source R software package *ltm* was chosen. Default settings such as the estimation method, which is marginal maximum likelihood (MML; Johnson, 2007) were used.

Second, simulation conditions were set based on conditions used in previous studies. Previous studies manipulated ability and item parameters as well as the number of items (*I*, i.e., test length), number of examinees (N), and number of repetitions (Abdel-fattah, 1994; Pan; 2012; Patsula & Gessaroli, 1995; Weiss & Von Minden, 2012; Yen, 1987). For this study, test length and number of examinees were varied while examinee abilities were randomly selected from a normal distribution. Item parameters were taken from an operational item pool from a grade eight statewide mathematics test. Item difficulty was approximately normally distributed in the original item pool. Item difficulty was then standardized and two random samples of 20 and 40 items were drawn from the item pool, which also showed approximately, normally distributed *b* values between -3 and +3 ($I = 40$, $\bar{x} = -0.01$, $s.d. = 1.57$; $I = 20$, $\bar{x} = -0.5$, $s.d. = 1.64$). Abilities were randomly generated to constitute an approximately normal distribution from -3 to +3, for 2000 examinees. Then subsets of 1000 and 250 abilities were randomly selected, which were also approximately normal. Finally, six simulation conditions were assembled (see Table 1). Dichotomous responses to items were randomly generated for each of these six conditions.

Table 1

Simulation Conditions

Condition Number	Number of Items (*I*)	Number of Examinees (*N*)
1	20	250
2	20	1000
3	20	2000
4	40	250
5	40	1000
6	40	2000

For generalizability of the results, the response generation process was repeated 100 times for each condition, which makes 600 generated response sets. These responses were then analyzed by *ltm* using 1PL, 2PL, and 3PL models to estimate item parameters and ability.

Third, the estimated item parameters and examinee abilities were compared with their respective true, simulated, values. Correlations, bias, mean absolute difference (MAD), and the root mean square error (RMSE) were calculated. Bias allows us to determine if deviations are greater in one direction than the other,

$$\text{Bias} = \frac{\sum_{j=1}^{J}\left(x_j - x_j\right)}{J}, \tag{1}$$

where x_j and x_j are the estimated and true values, respectively, and *j* represents the index over examinees or items, as appropriate, whether determining bias for item parameter or ability estimates. MAD is the average of the absolute value of the raw differences:

$$\text{MAD} = \frac{\sum_{j=1}^{J}\left|x_j - x_j\right|}{J} \tag{2}$$

The RMSE was used as a measure of overall error,

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{J}\left(x_j - x_j\right)^2}{J}} \quad (3)$$

Results

The comparisons were presented separately for 1PL, 2PL, and 3PL. Table 1 includes the results for 1PL, Table 2 and Table 3 includes 2PL and 3PL, respectively. Since estimates from the 3PL analyses with fewer than 1000 examinees were not stable and previous studies supported using 1000 examinees for 3PL (Lord, 1968; Yen, 1987), responses for only four conditions were analyzed with 3PL model.

Table 2

1PL Results

Condition		**Correlation**		**RMSE**		**MAD**	**Bias**
I	**N**	***b***	**θ**	**b**	**θ**	***b***	***b***
20	250	0.995	0.874	0.032	0.033	0.149	-0.007
20	1000	0.999	0.862	0.049	0.051	0.083	0.047
20	2000	0.999	0.863	0.012	0.006	0.049	0.006
40	250	0.994	0.933	0.030	0.032	0.148	-0.032
40	1000	0.999	0.927	0.052	0.051	0.081	0.052
40	2000	0.999	0.927	0.007	0.006	0.048	0.006

Table 3

2PL Results

Condition		Correlation			RMSE			MAD		Bias	
I	**N**	***a***	***b***	**θ**	***a***	***b***	**θ**	***a***	***b***	***a***	***b***
20	250	0.77	0.974	0.862	0.073	0.073	0.033	0.213	0.284	0.083	-0.062
20	1000	0.921	0.993	0.852	0.018	0.055	0.051	0.101	0.153	0.003	0.04
20	2000	0.958	0.997	0.855	0.012	0.018	0.006	0.071	0.104	0.003	0.001
40	250	0.772	0.976	0.928	0.076	0.05	0.032	0.197	0.252	0.091	-0.043
40	1000	0.926	0.994	0.921	0.011	0.044	0.051	0.091	0.136	0	0.044
40	2000	0.962	0.997	0.921	0.009	0.017	0.006	0.063	0.088	0.001	0.013

Table 4

3PL Results

Condition		Correlation				RMSE				MAD			Bias		
I	**N**	**a**	**b**	**c**	**θ**	**a**	**b**	**c**	**θ**	**a**	**b**	**c**	**a**	**b**	**c**
20	1000	0.573	0.9	0.142	0.758	0.155	0.234	0.057	0.207	0.306	0.493	0.131	0.174	0.241	0.061
20	2000	0.726	0.944	0.272	0.774	0.08	0.172	0.051	0.117	0.206	0.375	0.106	0.093	0.161	0.05
40	1000	0.547	0.91	0.367	0.831	0.127	0.168	0.04	0.22	0.28	0.45	0.11	0.148	0.09	0.04
40	2000	0.746	0.95	0.519	0.842	0.059	0.095	0.029	0.138	0.177	0.331	0.089	0.063	0.1	0.03

Discussion

Overall, the results indicated that *ltm* package gave accurate estimates with the 1PL, 2PL, and 3PL. Although 1000 examinees is suggested only for 3PL models, in all models accuracy increased with 1000 examinees and 40 items.

The difficutly (*b*) parameters and examinee abilities were estimated most accurately by *ltm*. However, estimating the guessing parameter was challenging. RMSE's reported in previous studies with 1000 examinees were between 0.11 and 0.15 for *a*, and between 0.10 and 0.14 for *b* (Gao & Chen, 2005; Kim, 2006; Yen, 1987). Therefore, RMSE calculations for *a* and *b* estimates were comparable with previous studies for the 1PL and 2PL. For the 3PL, estimating *b* values

with 1000 examinees and 20 items gave larger error. However, it should be noted that for the 3PL, it is suggested to use more than 1000 examinees especially with 40 items (Yen, 1987), particularly when using marginal maximum likelihood estimation, which is the default setting of *ltm*. Moreover, some 3PL analyses gave non-convergent solutions, as also observed in Lord (1968).

Some similarities were observed across all conditions. The accuracy increased in all three models as the number of examinees increased. A few exceptions to this statement are the 1PL results with 250 examinees and the 3PL with 1000 examinees where results were found to be almost the same between the 20 and 40 item conditions.

Educational Implications

The findings of this study, that IRT parameter estimates and examinee ability estimates found by *ltm* were comparable to true values, do have educational implications. Many researchers, especially, graduate students have been restricted by the IRT research they can conduct due to the cost of commercial software. This study can give researchers confidence to conduct IRT research using the open-source *ltm* package and will obtain comparable results to established software.

Future Work

It would be informative to replicate the analyses of this study using a commercial IRT software package, such as BILOG, then compare the accuracy of the results for the two packages. While this study focuses on dichotomous models, *ltm* is also capable of analyzing polytomous responses. A similar study to evaluate *ltm*'s item parameter and ability estimates in a polytomous context would be a natural extension of this work.

References

Abdel-fattah, A. (1994, April). Comparing BILOG and LOGIST estimates for normal, truncated normal, and beta ability eistributions. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading, MA: Addison-Wesley.

Bulut, O., & Zopluoglu, C. (2013, April). Item parameter recovery of the graded response model using the R package ltm: A Monte Carlo simulation study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18,* 351-380.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Sage Publications, Inc.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software, 20,* 1-24.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43,* 355-381.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Pan, T. (2012, April). Comparison of four maximum likelihood methods in estimating the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Patsula, L. N., & Gessaroli, M. E. (1995). A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Rizopoulos, D. (2013). Package "ltm": Latent trait models under IRT. http://rwiki.sciviews.org/doku.php?id=packages:cran:ltm

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling. *Journal of Statistical Software, 17*, 1-25.

Weiss, D. J., & Von Minden, S. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG*. Saint Paul, MN: Assessment Systems Corporation.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52, 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 [computer program]. Chicago, IL: Scientific Software.